Project Methodology: Malaysia Airlines Competitive Analysis

Executive Summary

This document outlines the methodology for analyzing Malaysia Airlines' competitive position against Qatar Airways, Singapore Airlines, and Emirates using statistical analysis and natural language processing techniques.

1. Project Context & Objectives

1.1 Problem Definition

Objective: Assess Malaysia Airlines' competitive positioning against top 3 global carriers to identify performance gaps and improvement opportunities.

Initial Questions: • Where does Malaysia Airlines lag behind competitors across service dimensions? • What are the statistically significant performance gaps? • How do customer sentiment patterns reveal operational strengths and weaknesses? • What opportunities exist for competitive repositioning?

1.2 Analytical Framework

Three-Stage Methodology:

1. Data Wrangling & Understanding: Quality assessment, standardization, exploratory analysis
2. Statistical Analysis: Hypothesis testing, effect size analysis, predictive modeling
3. NLP & Sentiment Analysis: Text mining, sentiment analysis, language pattern recognition

2. Data Foundation & Quality Assurance

2.1 Dataset Characteristics

• Source: Web-scraped airline review platform data • Temporal Scope: 2013-2025 (12-year period) • Volume: 8,137 reviews across target airlines • Coverage: Malaysia Airlines (1,471), Qatar Airways (2,600), Singapore Airlines (1,648), Emirates (2,418)

2.2 Data Quality Framework

Quality Assessment Strategy:

```
def assess_data_quality(df):
    # Missing data analysis with recovery strategies
    missing_analysis = calculate_missing_patterns()


    # Completeness scoring by airline
    airline_completeness = assess_airline_data_quality()


    # Quality categorization framework
    quality_scores = create_quality_scoring_system()
```

```
    return quality_framework
```

Data Quality Results: • Overall missing data: 26% → 2.3% after systematic recovery • High-quality data: 97.2% of final dataset • Cross-airline consistency validation completed

2.3 Data Standardization Protocols

Aircraft Data Standardization:

```
def standardize_aircraft(aircraft_str):
    # Boeing aircraft classification
    if 'boeing' in aircraft_str.lower():
        return categorize_boeing_variants()

    # Airbus aircraft classification
    elif 'airbus' in aircraft_str.lower():
        return categorize_airbus_variants()

    # Mixed fleet handling
    return handle_special_cases()
```

Route Categorization Framework: • Hub analysis: KL Hub vs Non-KL routes • Regional mapping: Asia, Europe, Middle East, Australia, Americas • Route type: Direct vs Connected flights

Temporal Analysis: • Period classification: Historical, Pre-COVID, Post-COVID • Seasonal analysis integration • Performance trend identification

3. Statistical Analysis Methodology

3.1 Competitive Benchmarking Framework

Descriptive Analysis:

```
def competitive_descriptive_analysis(df, airlines):
    # Service performance matrix calculation
    performance_matrix = df.groupby('airline')[service_cols].agg(['count', 'mean', 'std', 'median'])

    # Competitive gap analysis
    gap_analysis = calculate_performance_gaps()

    # Priority ranking system
    improvement_priorities = rank_improvement_areas()
```

```
    return competitive_summary
```

3.2 Statistical Testing Strategy

One-Way ANOVA Implementation:

```
def competitive_anova_analysis(df, airlines):
    anova_results = {}

    for service in service_dimensions:
        # Group data by airline
        groups = prepare_airline_groups(service)

        # Perform ANOVA
        f_stat, p_value = f_oneway(*groups)

        # Calculate effect size (eta-squared)
        eta_squared = calculate_effect_size(groups)

        anova_results[service] = {
            'f_statistic': f_stat,
            'p_value': p_value,
            'eta_squared': eta_squared,
            'significance': interpret_significance(p_value)
        }

    return anova_results
```

Effect Size Analysis:

```
def cohens_d_analysis(group1, group2):
    # Calculate pooled standard deviation
    pooled_std = calculate_pooled_std(group1, group2)

    # Cohen's d calculation
    effect_size = (mean(group1) - mean(group2)) / pooled_std

    # Practical significance interpretation
```

```
    interpretation = interpret_effect_size(effect_size)

    return effect_size, interpretation
```

3.3 Regression Modeling Approach

Service Impact Analysis:

```
def service_priority_regression(df):
    # Prepare regression variables
    service_predictors = ['seating_comfort', 'staff_service', 'food_quality',
                         'entertainment', 'value_for_money']

    # Standardized regression for coefficient comparison
    X_standardized = standardize_features(X)
    model = OLS(y, X_standardized).fit()

    # Business interpretation of coefficients
    importance_ranking = rank_service_importance(model)

    return model, importance_ranking
```

4. NLP Methodology

4.1 Text Preprocessing Pipeline

Advanced Text Cleaning:

```
def advanced_text_preprocessing(df):
    # Initialize NLP tools
    sia = SentimentIntensityAnalyzer()
    stop_words = create_enhanced_stopwords()

    # Text standardization
    df['review_clean'] = df['review'].apply(clean_text)

    # Bigram extraction
    df['bigrams'] = df['review_clean'].apply(extract_bigrams)

    # Sentiment scoring
```

```
    df['sentiment_score'] = df['review'].apply(lambda x: sia.polarity_scores(str(x))['compound'])

    return df
```

4.2 Bigram Network Analysis

Network Graph Construction:

```
def create_bigram_network_graph(df):
    # Collect bigrams with sentiment weighting
    bigram_sentiment = analyze_bigram_sentiment()
    bigram_frequency = count_bigram_frequency()

    # Network graph construction
    G = nx.Graph()
    for bigram in top_bigrams:
        words = bigram.split()
        G.add_edge(words[0], words[1], weight=frequency)

    # Sentiment-based node coloring
    node_colors = calculate_sentiment_colors()

    return G, visualization_data
```

4.3 TF-IDF Distinctiveness Analysis

Brand Positioning Intelligence:

```
def create_tfidf_analysis(df):
    # Prepare airline documents
    airline_documents = aggregate_reviews_by_airline()

    # TF-IDF vectorization
    vectorizer = TfidfVectorizer(
        max_features=1000,
        min_df=2,
        max_df=0.8,
        ngram_range=(1, 2),
        stop_words='english'
```

```
    )

    tfidf_matrix = vectorizer.fit_transform(documents)

    # Distinctive term identification
    distinctive_terms = identify_airline_distinctiveness()

    return tfidf_results
```

4.4 Aspect-Based Sentiment Analysis

Service Dimension Sentiment Mining:

```
def analyze_service_aspects_sentiment(df):
    # Define service aspect keywords
    service_aspects = {
        'Crew': ['crew', 'staff', 'attendant'],
        'Food': ['food', 'meal', 'dining'],
        'Seat': ['seat', 'comfort', 'legroom'],
        'Check-in': ['checkin', 'boarding', 'gate'],
        'Lounge': ['lounge', 'terminal', 'amenity'],
        'Refund': ['refund', 'compensation', 'cancel']
    }

    # Calculate aspect-specific sentiment
    for airline in airlines:
        for aspect, keywords in service_aspects.items():
            aspect_sentiment = calculate_aspect_sentiment(airline, keywords)

    return aspect_sentiment_matrix
```

5. Visualization & Analysis Output

5.1 Competitive Visualization Strategy

Multi-Dimensional Performance Representation: • Radar charts for service performance comparison • Gap analysis with directional indicators • Head-to-head competitive positioning • Temporal trend analysis with recovery patterns

5.2 NLP Visualization Framework

Text Mining Visual Intelligence: • Bigram network graphs with sentiment coloring • TF-IDF importance charts with sentiment weighting • Multi-quadrant word clouds (positive/negative/competitor excellence) • Service aspect sentiment comparative analysis

6. Statistical Validation & Quality Assurance

6.1 Reproducibility Framework

Systematic Validation Approach:

```
# Set reproducible random states
np.random.seed(42)
random.seed(42)

# Consistent data processing
df_processed = apply_consistent_preprocessing()

# Cross-validation of statistical tests
validate_statistical_assumptions()

# Effect size interpretation standards
apply_cohen_guidelines()
```

6.2 Cross-Method Validation

Convergent Validity Assessment: • Statistical gaps validated against sentiment gaps • ANOVA significance confirmed through effect sizes • Regression coefficients aligned with aspect sentiment analysis • Quantitative findings supported by qualitative language patterns

7. Implementation Framework

7.1 Priority Matrix

Evidence-Based Recommendation Framework:

1. Statistical Significance: All service gaps tested at $p<0.05$ level
2. Practical Significance: Cohen's d effect size interpretation
3. Regression Impact: Standardized coefficient ranking for resource allocation
4. Sentiment Validation: Customer language pattern confirmation

7.2 Implementation Pathway

Phased Improvement Strategy: • Phase 1 (0-6 months): Address top 3 statistical priority areas • Phase 2 (6-12 months): Implement high-impact regression targets • Phase 3 (12-24 months): Strategic positioning against Qatar Airways

8. Limitations & Methodological Considerations

8.1 Data Limitations

• Review platform selection bias • English-language limitation • Temporal COVID-19 effects • Missing operational cost data

8.2 Analytical Constraints

• Cross-sectional analysis limitation • Sentiment analysis tool limitations • TF-IDF parameter sensitivity • Effect size interpretation subjectivity

9. Quality Assurance & Validation

9.1 Statistical Rigor

• Multiple testing correction consideration • Effect size practical significance • Regression assumption validation • Cross-validation methodology

9.2 NLP Validation

• Sentiment analysis tool validation • Bigram network meaningful connection filtering • TF-IDF parameter optimization • Aspect keyword validation through domain expertise

---

Conclusion

This methodology applies data science techniques to competitive intelligence analysis through a three-stage approach providing quantitative statistical analysis validated by qualitative sentiment intelligence. The framework ensures reproducible results while maintaining relevance through clear improvement prioritization.

Key Methodological Components: • Statistical testing with effect size interpretation • Advanced NLP techniques with network analysis • Cross-method validation for result confidence • Business intelligence translation • Reproducible analytical pipeline

The methodology establishes a framework for ongoing competitive intelligence monitoring and strategic decision-making in the airline industry.